We sincerely thank the reviewers for providing thoughtful comments on our manuscript. We have provided a point-by-point response to these comments below. We are confident that these additions and revisions improve our manuscript for readers of Science of the Total Environment.

In addition to making revisions that address points raised by the reviewers, we have also taken the opportunity to make minor improvements that we identified.

## Reviewer 1

This manuscript is a welcome addition in the world of trend analyses. The authors propose combinations of GAM and mixed-effects meta-analysis regression to account for uncertainty. The methodology is sound and results are convincing. My recommendation would be even more enthusiastic if the authors had formatted their methods sections in a more traditional manner. I am of the opinion that models should be presented using clear mathematical equations and not in R code semantics. This comments and a few other can be found in the attached document.

* **Response**: We really appreciate these comments that affirm our opinion on the value of this work for trend analysis of long-term water quality data. We have changed our formulas and text to a conventional format, as opposed to R code notation. Briefly, here are the updated equations for the models described in the methods. All other instances in the main text referring to R code (except in a few locations) have been changed. Also please note that the attached pdf that was included with the reviewer’s response did not include any markup comments. This is contrary to what the reviewer indicated.

## Reviewer 2

### Overview

This manuscript presents a new application of GAMs approaches to fit irregular time series data (gaps, changing frequency of measurements) to estimate trends. The paper presents a case study using Chl a data from San Francisco Bay. The manuscript is presented as a general approach with applications to other similar environmental time series data. A comparison of methods shows that the GAMs plus meta-analysis method provides some different trend conclusions from ordinary least squares and plain GAMs. The paper is generally well written and well-suited for the audience for STOTEN. I have some clarifications that would help improve/clarify the utility and readability of the manuscript. These are presented below.

* **Response**: Thank you for your comments on our manuscript. We have addressed your concerns on the utility and readability of our manuscript below.

### Utility (1):

We applied the wqtrends code in R to a time series data set of X2 (salinity intrusion measure) over approximately in San Francisco Bay. Some observations:

-the code did not work with the full time series of daily X2 data. But we were able to make it work using a random sampling of 5 points per month. Could the reason for this error be explored? As time series data are concerned, we provided an input of ~100 years of daily data. A sample result for the trends are shown in the figure below. I think it would be helpful for readers/users to understand what the practical limitations of the code/algorithm are.

* **Response**: Very rarely do reviewers actually apply the proposed methods on novel datasets as a proof of concept. For that, we are very appreciative of your efforts in exploring the R package. These applications are important for testing both the method and software. As such, you exposed an important limitation of our method that was not clearly articulated in the draft. We provide two responses to this limitation.
* First, the X2 dataset is at the daily scale, so one aspect of the GAM smoothing methods proposed in our manuscript for “irregularly spaced or missing” monitoring data is unnecessary. The GAMs provide two benefits: smoothing irregularly spaced data and estimating the non-independence of sequential data so that propagated uncertainty does not incorrectly assume sequential data are independent. The modelling approach is designed to provide both a continuous (e.g., daily) estimate of the long-term trend and a measure of uncertainty in the absence of more regularly collected values. So, application of our method to datasets with sampling intervals longer than the daily scale would be the most compelling use case, but they are nevertheless relevant for uncertainty propagation even with daily data. However, because some of the daily X2 values are in fact imputed, it would be appropriate to use the raw data only in our method, and possibly to enrich the GAM with whatever domain-specific model is used for smoothing X2 values, in order to fully propagate uncertainty.
* Second, and more important, the practical limitations of applying a GAM with flexible smoothing to nearly 100 years of daily data is beyond the computational power of most desktop computers. This is asking too much of the smoothing splines used in the mgcv package. Given this limitation, we do feel it is important to clearly state the intended use case for these models to avoid confusion in application. We have added clarification both in the manuscript and the package vignette to clearly indicate the type of data that is appropriate for the methods. In practice, if one wants to analyze such a long series, one could break it down into multiple (possibly overlapping) shorter series for the first two stages (GAM estimation and seasonal averaging with uncertainty propagation). The only difference would be that the estimated optimal degree of smoothness might change across different time windows, but that may be appropriate and supported by having so much data.
* In the manuscript, lines 194 - 196: “Approximate monthly or biweekly sampling with coverage of at least a decade is common for many long-term monitoring programs and is the motivating use case for the methods herein.”
* In the package vignette, first paragraph (<https://tbep-tech.github.io/wqtrends/articles/introduction.html>): “These models are appropriate for data typically from surface water quality monitoring programs at roughly monthly or biweekly collection intervals, covering at least a decade of observations (e.g., Cloern and Schraga 2016). Daily or continuous monitoring data covering many years are not appropriate for these methods, due to computational limitations and a goal of the analysis to estimate long-term, continuous trends from irregular or discontinuous sampling.”
* We have also added a paragraph in the discussion regarding limitations of the approach, one of which is related to the issues above:
* Lines 532 - 541: “Several limitations of the proposed methods deserve mention. First, if sampling is so irregular that important fluctuations are missed entirely in some years, the GAM estimates and uncertainty propagation could become dubious in interpretation and usefulness. Second, estimation of GAMs for very long series can be computationally demanding. When this is an obstacle, one could do the first two analysis stages using temporal windows of the full data, with the only implication being that different degrees of smoothness may be estimated for different windows, which indeed might be justified by the data. Third, meta-analysis regression results for a very small number of years, particularly confidence intervals and associated p-values, may be inaccurate (e.g., in confidence interval coverage). In such cases, one could make alternative use of the GAM seasonal averages and standard errors, such as for pairwise comparisons among years.”

### Utility (2):

From figure 7 it seems that there are very different conclusions for Station 36 depending on the method used. From Figure 8, however, it seems that Station 36 is a bit of an anomaly in this regard, and that other 8 stations in the case study have findings on trends that are similar among the three methods—and that the results are not dramatically different. Perhaps this should be the basis for added discussion in the final section—when is the added complexity of the GAMS and meta analysis method is appropriate in a real-world setting? Alternatively, it would be helpful if the authors were to examine another data set from the region that they work with, to review the relative findings of trend significance and what this depends on.

* **Response**: You are correct to note that many of the comparisons had similar results and we agree that it is useful to clarify that our approach is warranted relative to more conventional methods. If a simpler approach is used, there is no way to know for sure that an invalid result is obtained without applying our proposed methods. As such, we argue that using our approach is still the wiser option in all cases. Otherwise, a user cannot be certain the correct conclusions are obtained using simpler methods. The paragraph on lines 498 to 513 was revised to address these points:
* “Incorrect conclusions on trends can have dramatic consequences for regulated parties under existing water quality compliance frameworks (Smith et al., 2001). Our examples in Figures 7 and 8 demonstrate these risks if propagation of uncertainty from raw observations across methods is unaccounted for in trend assessment. The ‘naïve’ method using OLS regression applied to seasonal averages from the raw observations fails to propagate uncertainty, similarly to averaging results within a year and applying a simple Kendall test. In some cases the results may be similar to those from fully propagating uncertainty, but the loss of information can lead to increased Type I or II error rates depending on characteristics of the raw data and the method used for their evaluation (Shabman and Smith, 2003). Our examples demonstrated the increased potential for incorrect conclusions at specific monitoring locations and, at larger spatial scales, across all stations if simpler trend analyses are used. Even though simpler methods may produce similar results in some cases, particularly with frequent sampling and similar effort between years, the only way to confirm such an outcome would be to compare results, relying on the method with full propagation of uncertainty to be the more robust method. Use of methods that fully account for uncertainty is recommended to obtain statistically valid results in a wider range of conditions.”
* We also want to mention that we recently discovered a minor bug in the seasonal feature extraction of wqtrends that prevented an estimate for years on the tail ends of the time series with incomplete data. Because of this, we have updated Figure 7 and 8, now using stations 30 and 34 as examples in the manuscript. None of the conclusions change in the manuscript.

### Readability/Clarity:

The term meta-analysis seems to be an important part of the paper and is referred to multiple times, but it is never clearly explained in the manuscript. For example, at line 164. See the following sentences.

“Third, we used a mixed-effects meta-analysis to estimate trends and test hypotheses about the change in seasonal averages across years. While 166 meta-analysis methods arose from analyses of results from multiple studies, their distinguishing characteristic is propagation of uncertainty (Gasparrini et al., 2012; Sera et al., 2019). Meta-analysis uses response data that includes standard errors (uncertainties) as needed to address our questions.”

It is not clear to me what exactly is meant by meta-analysis and what is being tested in the context of the time series data. An explanation that is understandable by a general audience of potential users of this method would be helpful. It does seem likely, as that sentence points out, that most readers would be more familiar with meta-analysis methods in the context of aggregating the results of multiple studies. There are other helpful citations afterwards, but a more explicit connection between the familiar use case and the use case here of combining different seasonal features from a single regression model would have been helpful for reading this work.

* **Response**: Thank you for pointing out this ambiguity. We agree that a distinction between the more common understanding of “meta-analysis” and the one applied here needs to be made early in the manuscript. We have added content in the introduction to clarify our methods.
* Lines 125 - 135: “Meta-analysis regression incorporates a known (or estimated) standard error for each response datum. Usually meta-analysis is used when each response summarizes a dataset from a separate study, along with its standard error, and the meta-analysis looks for across-study patterns in effect size (i.e., Lortie et al. 2014). Here each response summarizes one year or season of data, with standard error from the GAM, and the meta-analysis looks for patterns across years. Thus, while meta-analysis methods are most commonly associated with combining results from multiple studies into a larger analysis, their key modeling step is propagation of uncertainty (Gasparrini et al., 2012; Sera et al., 2019). To do this, meta-analysis makes use of a known (estimated) standard error for each response datum, which is the required priority here to propagate standard errors from the GAM into a regression of seasonal averages.”